Token Activation Map to Visually Explain Multimodal LLMs

Supplementary Material

Section and Textual Content	**Graphic and Tabular Content**
Supp. A: Tabular Catalog of the Supplementary.	Table 5: The **catalog** for quick reference.
Supp. B: Examples of Multimodal Activation Map	Fig. 9: **How to read** these visualization examples?
Supp. C: TAM to Explain All Generated Tokens.	Fig. 10: A **complete example** explaining all tokens.
Supp. D: Implementation of Baselines.	N/A, MLLM and explainability **baselines**.
Supp. E: Details of Metrics.	N/A, Invalid faithfulness evaluation for MLLM and other metric details.
Supp. F: Analysis About Motivation.	N/A, Statistical test and causal validation.
Supp. G: Extensive Cases About Method Comparison.	Fig. 12 & Fig. 13: TAM exceeds existing **SoTA methods** in extensive visualizations.
Supp. H: Visual Comparison About Causal Inference	Fig. 14: How does the **estimated causal inference** work in visualization?
Supp. I: Visual Comparison Among Denoise Filters	Fig. 15: How does the **rank Gaussian filter** work in visualization?
Supp. J: Visualization of Ablation Study.	Fig. 16: Two involved modules are **mutually beneficial**.
Supp. K: Explainability Results on Diverse MLLMs.	Table 6: MLLM **quantitative results**; Fig. 11 relation between **model size** and explainability.
Supp. L: TAM for MLLM Visual Comparison.	Fig. 17 & Fig. 18: TAM supports visual **comparison among MLLMs** about attributes.
Supp. M: Extensive Cases About Attributes analysis.	Fig. 20 & Fig. 21 & Fig. 22 & Fig. 23: Explaining **fine-grained attributes** beyond SoTA.
Supp. N: TAM for Biased Scenario.	Fig. 19: TAM supports **biased error analysis**.
Supp. O: Extensive Failure Cases Study.	Fig. 24 & Fig. 25: TAM supports **failure cases analysis** for images and videos.
Supp. P: Extensive Success VQA Examples.	Fig. 26: Explanation result on the **VQA** dataset.
Supp. Q: Examples About Video Visualization.	Fig. 27: Clearer **video visualizations** with fewer redundant activations and noises.
Supp. R: Corner Case About Reasoning.	Fig. 28: TAM supports failure case analysis for **visual reasoning**.
Supp. S: TAM for Multi-image Conversation.	Fig. 29: High applicability on **multi-image** conversation.
Supp. T: TAM for Multi-turn Conversation.	Fig. 30 & Fig. 31: TAM supports **multi-turn conversation** about attributes and case study.

Table 5. Tabular catalog of the supplementary.

A. Tabular Catalog of the Supplementary

In this supplementary material, we primarily provide extensive qualitative results to demonstrate the effectiveness and wide applicability of TAM. These sections include comparisons with state-of-the-art (SoTA) methods, visualizations about ablation study, attribute explanation, failure case analysis, VQA examples, video visualizations, MLLM comparisons, reasoning analysis, multi-turn conversation, multi-image input, as well as some quantitative results and baseline descriptions. To enhance readability given the extensive content, we provide a tabular catalog in Table 5 for quick reference.

B. Examples of Multimodal Activation Map

In this section, we present a high-resolution example accompanied by detailed captions to facilitate the explanation of the multimodal activation map defined in Eq. 3. The primary element is the activation map at the top, which reflects the degree of vision-text alignment and serves to visually explain the MLLM. All multimodal activation maps in this paper adhere to a consistent format, and we provide high-quality images; please zoom in if any example appears too small to read.

The visual activations and textual relevances are normalized to the same scale as specified in Eq. 3, allowing for a direct comparison between the two modalities to identify where the model focuses—whether on the image or the context. The text is colored by tokens, with some words represented by multiple tokens marked in different colors. The answers following the target are not visible for the current explained token, and are therefore colored in gray. The colors of the candidate responses reflect the prediction confidence of the top three tokens corresponding to the target, which can be useful for analyzing failure cases through potential predictions and confidence levels associated with each token.

C. TAM to Explain All Generated Tokens

The proposed TAM demonstrates a significant advantage in explaining multiple generated tokens from MLLMs, in contrast to conventional models that typically focus on a single output. We depict all multimodal activation maps in Fig. 10 and support it. The figure clearly shows that TAM produces considerably fewer redundant activations, particularly for non-object words, thanks to the proposed estimated causal inference.

For instance, the activations for the function word "with" and the punctuation mark "." in the Class Activation Map (CAM) are extremely high, overshadowing object activations. Additionally, these visual activations often exceed those of text tokens, resulting in excessive redundant activations. In comparison, our TAM is much clearer and focuses more on important objects. These results suggest that TAM produces closer explanations to the understanding of humans than CAM, where words from the image are highlighted while those related to texts show much fewer responses. Besides, there are much fewer activations showing higher visualization quality as well.

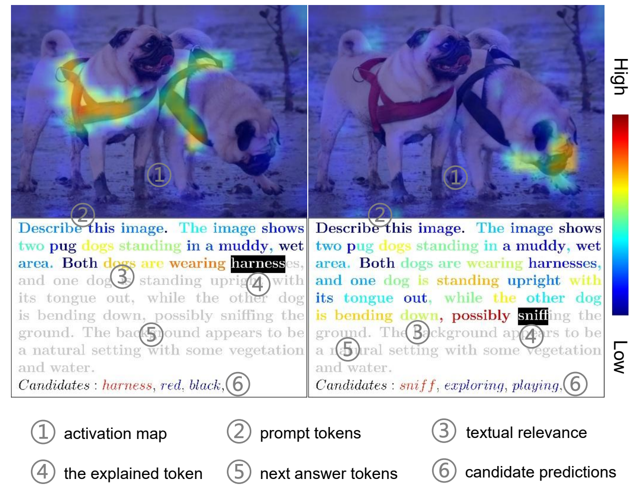


Figure 9. A high-resolution example of the multimodal activation map. This image is processed by the Qwen2-VL-2B model [51]. There are five components to visualize, including the visual activation map, prompt tokens, textual relevance, the explained target token, next answer tokens, and its top predictions (top 3). The colors indicate the corresponding response degree.

D. Implementation of Baselines

We conduct experiments on various MLLMs, including Qwen2-VL-2B [51], Qwen2-VL-7B, LLaVA1_5-7B [35], LLaVA1_5-13B, InternVL2_5-2B [15], InternVL2_5-4B, and InternVL2_5-8B implemented by transformers using weights from huggingface. For Qwen2-VL all the models the weights are the Instruct version (e.g., Qwen2-VL-2B from huggingface "Qwen/Qwen2-VL-2B-Instruct"). The example of LLaVA1_5 weights is from "llava-hf/llava-1.5-7b-hf" and "thisisiron/InternVL2_5-2B" for InterVL2_5. Due to device limitations, very large MLLMs are not used. For image resolution, Qwen2-VL supports raw image size, while LLaVA1_5 and InternVL2_5 fix image sides at 336 and 448, respectively. For the implementation of video caption on Qwen2-VL [51], we extract 10 frames from a short video and repeat frames for the number of temporal_patch_size to ensure each frame has its own activation, instead of activations from other frames. We use the same prompts for the involved MLLMs. These prompts are set according to the average length of captions. For COCO Caption [13] the prompt is "Write a one-sentence caption for this image:", and the prompts for GranDf [42] and

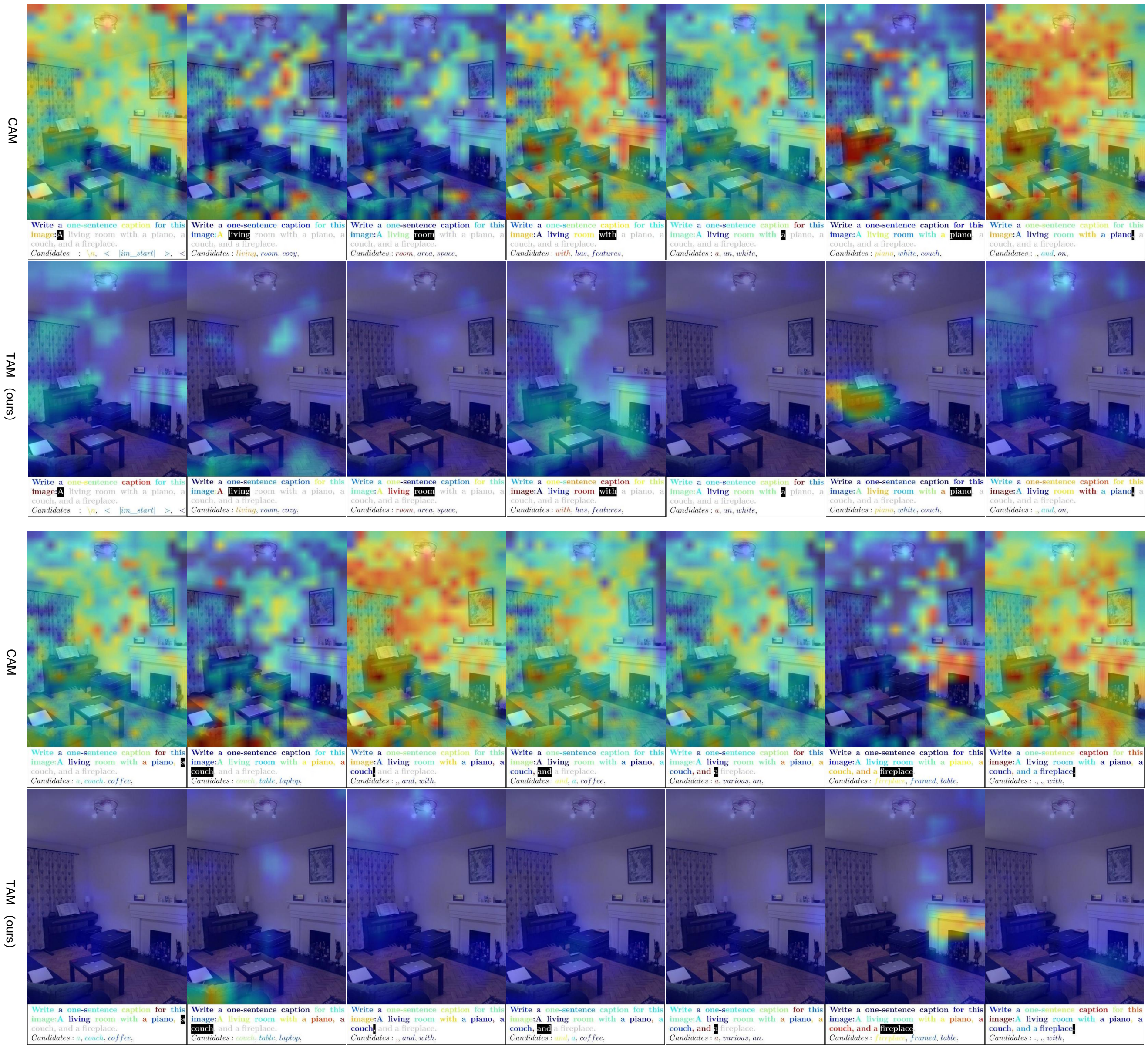


Figure 10. Visualization of one example for all generated tokens on the Qwen2-VL-2B [51] model. *The proposed TAM shows more accurate object localization ability beyond conventional CAM, with much fewer redundant activations in an overall view*.

OpenPSG [57] are "Write a description for this image using around two sentences:", "Write a description for this image using around three sentences:", respectively. For attribute analysis, the prompt is set like "What is the [attribute] in this picture?". In addition, we use the prompts provided from the QK-VQA [37] and STAR [52] datasets, which vary according to the images.

For the explanation baselines [2, 9, 26, 32, 45], we implement them referring to their official codebases. To obtain attention weights for attention-based explainability methods (such as attention weights and Rollout [1]), we rewrite the SdpaAttention in PyTorch. This is necessary because the original implementations of SdpaAttention and FlashAttention do not provide attention weights. For the CP-LRP [5] and AttnLRP [2], modules of MLLM are replaced by the official implementations of AttnLRP to back-propagate the relevance from output to visual tokens. Besides, these methods need to close the kv_cache to maintain gradients for key and value. Note, the Grad-CAM [45] is equivalent to CAM. Because the weights for the activation map are derived from the classifier weight in MLLMs. Since there is only a feature vector without the pooling and other structures. The gradient is fully dependent on the classifier weight at the same ratio among channels. Since they are the same, we apply the classifier weight of CAM to achive TAM to avoid extra back-propagation in Grad-CAM. Other implementation methods of estimated causal inference include the mean of context maps (mean), using the attention weights as relevance in Eq. 4 (AttnWeights).

E. Details of Metrics.

The metrics are based on the part of speech using the pos_tag function from the NLTK Python package. The specific tags are "NN", "NNS", "NNP", and "NNPS" for Obj-IoU. Function words are identified by the tags: "CC", "DT", "EX", "MD", "POS", "PRP", "PRP$", "UH", "WDT", "WP", "WP$", and "WRB". Notably, we exclude the tags "IN" and "CD" from function words, as they pertain to location and quantity.

Besides the IoU based metric, in Sec. 3.4, we have discussed the difference between the used plausibility test (how accurately it reflects the true reasoning process), compared with another widely used faithfulness (how accurately it reflects the true reasoning process). While the faithfulness is not suitable for MLLM evaluation. Because the perturbation tests [2, 11] of the faithfulness metric alter the generated texts every time, resulting in inconsistent generated texts that are not stable to evaluate raw generated texts of MLLM. Specifically, masking different input regions in the faithfulness test drastically changes MLLM output tokens, invalidating observations of "decision-making" tied to a fixed class. In conventional models, input changes affect a fixed class's confidence, but in MLLMs, it causes vanished tokens or shifted context, making confidence comparisons invalid. Besides, its cost is unacceptable, which needs N times repeated inferences (N = token number $\times$ regions ratios).

Another consideration regarding the metrics is the variation in response levels. Specifically, our ECI involves a subtraction operation between activation maps, which can lead to a lower overall intensity compared to the original responses. We did not overlook this limitation when design the metrics; instead, we implemented a straightforward operation to penalize excessive discrepancies. Specifically, we use the response map of the first prompt token in place of the first generated token in evaluation. Since the first prompt token does not have any earlier text tokens, the map does not incorporate the ECI and reflects the original response level. If the response level of the altered map significantly differs from the processed maps, it can result in inappropriate background thresholds, thereby diminishing the Func-IoU metric. For instance, if the background threshold processed after ECI is 0.1, it may be too low for this map, leading to false positives and consequently affecting the metrics. Detailed operations can be referenced in our open-source code.

F. Analysis About Motivation

In Fig. 1(c), we randomly pair CAMs and count their L1 distance against text correlation. Higher text correlation corresponds to lower distance, indicating concurrent interferences. In this section, we provide a statistical test to support it. Specifically, the added statistical test is the Pearson correlation at -0.16 with p-value of 1.5E-30. Since most pairs are not related in the random pairing, the correlation is not strong. When pairing the most related tokens, the Pearson correlation comes to -0.359 (p-value 7.9E-32). It confirms that the negative correlation is evident.

We also conducted a causal validation for the causal inference. In this paper, our ECI is based on the potential outcome model (POM). The used causal validation for this model is the Placebo test. Specifically, we validate it by replacing the target CAM to a random earlier CAM as the placebo (not the observed target), and then record the results drop. The Obj-IoU reduced to 6.2% on COCO Caption and 4.4 times lower than the raw result, suggesting the causal effect is significant.

G. Extensive Cases about Method Comparison

In addition to the visual comparison presented in Fig. 2, we offer more complex examples in Fig. 12 and Fig. 13 within this section. The findings are consistent with those discussed in Sec. 4.3: the proposed TAM significantly outperforms existing explainability methods.

Specifically, TAM generates fewer redundant activations and exhibits less noise compared to gradient-based methods [9, 45]. Moreover, it effectively locates objects, contrasting with the scattered activations seen in attention-based methods (e.g., Attention, Attention-Rollout [1], CP-LRP [5], AttnLRP [2]). These results indicate that TAM enhances the localization capabilities of MLLMs, even in complex scenarios. Consequently, TAM can be integrated into existing MLLMs without requiring grounding abilities, thereby facilitating a wide range of potential downstream tasks without additional supervision or alignment.

H. Visual Comparison About Causal Inference

We have validated the effectiveness of the proposed Estimated Causal Inference (ECI) in Table 2. In this section, we present visualization results that illustrate how our ECI outperforms existing methods and alternative implementations, as shown in Fig. 14. The first baseline we consider is feature surgery [32], which is designed to mitigate redundant features along the class dimension. However, the challenge with multi-language models (MLLMs) lies in the correlated activations along the token prediction dimension, which is fundamentally different. As a result, feature surgery performs significantly worse than our ECI. Given the limited methods addressing correlated activations, we introduce additional baselines derived from other implementations of ECI: ECI-mean and ECI-attnWeights, as details in Supp. D. Although these suboptimal implementations outperform feature surgery, they still yield inferior results compared to the final ECI. Notably, our ECI demonstrates superior performance in handling function words, producing significantly fewer redundant activations while achieving better recall of target objects. These results indicate that our ECI is well-designed and effective for mitigating correlated activations among the generated tokens of MLLMs.

I. Visual Comparison Among Denoise Filters

Image denoising remains a traditional research topic, but it is the first time to be introduced in the visual explanation field. The issue of noise has been addressed in Sec. 2, where various methods aimed at noise reduction in transformers are discussed. However, residual noise persists even after these methods are applied. Consequently, it is essential to introduce denoising filters as a straightforward yet effective solution. Unlike conventional models that produce very small output sizes (e.g., 7×7), the output size of MLLMs is comparatively larger (e.g., 36×36). As a result, scatter-shaped noise is more likely to occur in MLLMs.

These noises belong to the salt-and-pepper noises in general, which can be effectively addressed using median and Gaussian filters. While these methods do not represent the optimal solution, as illustrated in Fig. 15. Specifically, the Gaussian filter proves inadequate in mitigating clustered noise, leaving many noises visible in the yellow boxes. The median filter reduces noise effectively, yet it still leaves behind unsolved scatter noise, with additional missing regions indicated by blue boxes in the final row. Similarly, the adaptive median filter [8] exhibits significant scatter noise, particularly near image edges where noise concentration is higher. In contrast, our proposed rank Gaussian filter demonstrates superior performance by amalgamating the strengths of both Gaussian and median filters, along with the novel technical enhancements discussed in Sec. 2.

J. Visualization of Ablation Study

We conducted ablation studies in Table 1. In addition to the quantitative results, we present further visualizations in Fig. 16 to elucidate the effectiveness of these modules. The first column showcases the baseline method, CAM [56] / Grad-CAM [45], which displays numerous redundant activations accompanied by noise, highlighted in white boxes. The proposed estimated causal inference (ECI) method in the second column effectively mitigates most correlated activations, although some persistent noise remains. The rank Gaussian filter in the third column successfully removes this noise, but redundant activations are still evident. By integrating these two innovative techniques into the proposed TAM, we achieve substantial explanatory results that leverage the strengths of both approaches. These examples illustrate the mutual benefits of the modules, leading to an overall improvement that exceeds the sum of their individual contributions, as shown in Table 1.

K. Explainability Results on Diverse MLLMs

In addition to the explainability improvements highlighted in Table 4, we present specific results in Table 6. This table reveals that the overall F1-IoU of the TAM ranges from 30.68% to 41.45% across three datasets and seven MLLMs. In contrast, the baseline CAM [56], which is considered a SoTA method in terms of performance and practicality (as shown in Table 3), achieves F1-IoU results ranging from 23.63% to 34.39%. These results clearly indicate that TAM demonstrates broader applicability and enhanced explainability across diverse MLLMs.

Moreover, TAM offers a unique perspective on evaluating MLLMs from an explainability standpoint, beyond existing metrics. For instance, LLaVA models [35] and Qwen2-VL models [51] exhibit higher F1-IoU scores than InternVL models [16] on the COCO Caption [13] and OpenPSG [57] datasets, while InternVL models excel on the GranDf dataset [42].

We also observe scalability in explainability across certain model sizes, as shown in Fig. 11a-c. For example, LLaVA models show improvements from 7B to 13B, InternVL models from 2B to 4B, and Qwen2-VL models from 2B to 7B. This trend suggests a positive correlation between the scalability and explainability of MLLMs to a certain extent. When the model size are larger, the model tends to encode objects with fewer tokens, leading to a decrease in recall (see Fig. 11d) and an increase in precision. Subsequently, the Obj-IoU decreases due to a more significant decrease in recall.

L. TAM for MLLM Visual Comparison

Generally, researchers MLLMs using quantitative metrics or textual outputs, while visual comparisons remain underexplored. In contrast, visual evaluations are prevalent in conventional models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), which effectively illustrate the enhanced representational capabilities of new models. The absence of a dedicated explanation tool for MLLMs may contribute to this gap. The proposed TAM addresses this limitation, enabling researchers to conduct visual comparisons of their MLLMs against existing models, beyond the conventional textual comparison.

We present visual comparisons among MLLMs in Fig.

Method	MLLM	COCO Caption			GranDf			OpenPSG		
		Obj-IoU	Func-IoU	**F1-IoU**	Obj-IoU	Func-IoU	**F1-IoU**	Obj-IoU	Func-IoU	**F1-IoU**
CAM	LLaVA1_5-7B [35]	23.17	43.16	30.16	20.07	47.48	28.21	25.11	51.55	33.77
TAM		27.65	61.43	**38.13**	20.71	59.15	**30.68**	28.57	61.06	**38.93**
CAM	LLaVA1_5-13B [35]	24.82	51.18	33.43	21.34	43.99	28.74	26.65	48.45	34.39
TAM		29.12	58.5	**38.88**	22.1	51.02	**30.84**	30.88	59.96	**40.76**
CAM	InternVL2_5-2B [16]	15.94	45.62	23.63	18.28	37.64	24.61	19.76	46.42	27.72
TAM		21.38	65.1	**32.19**	20.48	85.93	**33.08**	23.0	86.86	**36.36**
CAM	InternVL2_5-4B [16]	18.23	40.95	25.23	20.91	44.52	28.46	21.28	34.7	26.38
TAM		21.76	63.12	**32.36**	22.53	89.71	**36.02**	23.49	89.75	**37.23**
CAM	InternVL2_5-8B [16]	14.59	64.41	23.8	18.04	57.42	27.45	18.46	62.21	28.47
TAM		19.98	66.53	**30.73**	21.56	85.95	**34.47**	21.73	88.74	**34.91**
CAM	Qwen2-VL-2B [51]	21.23	51.93	30.14	17.85	62.15	27.74	22.93	48.5	31.15
TAM		27.37	68.44	**39.1**	18.65	88.97	**30.83**	26.26	92.99	**40.95**
CAM	Qwen2-VL-7B [51]	22.51	42.44	29.42	18.6	68.03	29.21	23.41	42.94	30.3
TAM		28.13	71.85	**40.43**	19.88	90.57	**32.61**	26.94	89.88	**41.45**

Table 6. TAM shows wide applicability on diverse MLLMs and datasets beyond the CAM [56] for all the experiments on the major F1-IoU (%) metric at large margins. TAM can be used as a visual comparison approach, where Qwen2-VL models [51] show better visual explainability than LLaVA1_6 [35] and InternVL2_5 [16] on the COCO Caption [13] and OpenPSG [57] datasets.

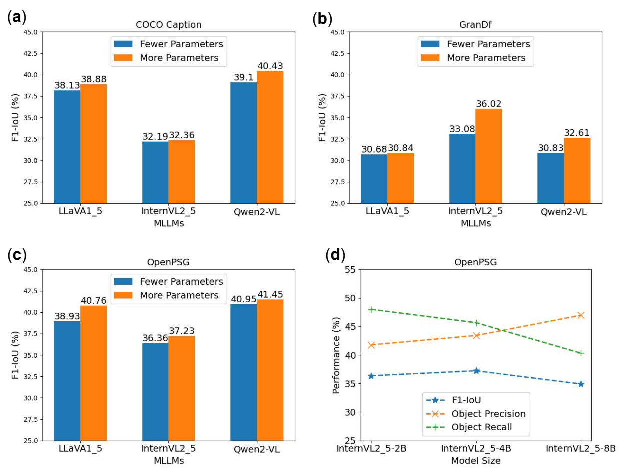


Figure 11. Relation between model size and explainability. (**a**-**c**) F1-IoU (%) results on LLaVA1_5 [35] (7B, 13B), InternVL2_5 [16] (2B, 4B), and Qwen2-VL [51] (2B, 7B) across diverse datasets [13, 42, 57] indicates the explainability is increased with more model parameters within a certain range. (**d**) Increasing the parameters of InternVL2_5 [16] on the OpenPSG dataset [57] improves object precision; however, this comes at the cost of decreased recall, which may negatively impact the F1-IoU score when the recall is too low.

17, focusing on attributes, and Fig. 18, which highlights object recognition. One key finding is that Qwen2-VL-7B [51] surpasses LLaVA1_5-7B [35] in text attributes, as illustrated in Fig. 17. Furthermore, it demonstrates superior performance in the object cases compared with LLaVA. For instance, Qwen2-VL-7B exhibits stronger activations in the top two rows of Fig. 18 and lacks correlation with the sandwich when interpreting the pizza in the third row.

In contrast, InternVL2_5-8B [16] tends to overemphasize textual information, resulting in weaker visual responses. Although it generates highly detailed textual content that exceeds the length of LLaVA's outputs, its activation map displays comparatively lower intensity. This is evidenced by the pronounced responses in textual areas (illustrated by the increased red regions). Despite reasonable activations in several successful cases (indicated by the light blue regions), InternVL2_5-8B exhibits lower quantitative performance, as shown in Table 6.

These visual findings underscore that TAM offers a novel perspective for comparing MLLMs, providing deeper insights, particularly when quantitative results are closely matched across certain datasets. Researchers can leverage visual comparisons to highlight the advantages of their models.

M. Extensive Cases about Attributes Analysis

The proposed TAM enables users to analyze the fine-grained attributes of MLLMs. These attributes contribute to a deeper understanding of how the model works. We present various visualizations of attributes, including actions and colors in Fig. 20, text and shapes in Fig. 21, and locations for both images and videos in Fig. 22.

The results indicate that the tested model, Qwen2-VL-7B [51], possesses the capability to comprehend diverse attributes with a high degree of explainability. Furthermore, we compare existing methods [45, 56] with our proposed TAM in Fig. 23, where our method demonstrates significantly superior explanation quality. These activation maps provide visual evidence for the generated content, thereby enhancing the model's credibility.

N. TAM for Biased Scenario

The Task-Aware Mask (TAM) framework is capable of supporting the analysis of biased scenarios. In Fig. 19, we investigate whether the background environment unexpectedly influences the classification of target categories. The output text indicates that images of terrestrial birds with synthetic aquatic backgrounds were misclassified as waterbirds, suggesting a significant bias introduced by background features in the model's predictions. We conducted an in-depth analysis of this phenomenon using the TAM.

TAM effectively separates the contributions of different regions within the image to the classification decision, allowing for precise localization of the source of bias. Our research reveals that the synthetic aquatic background exerts a substantial influence on the model's internal representations, leading it to favor categorizing images as waterbirds. This finding underscores the importance of considering background information during the model training and evaluation processes. Over-reliance on background features rather than the characteristics of the target itself may result in systematic misjudgments in scenarios that include synthetic or artificially manipulated backgrounds. The TAM-based analysis provides an effective diagnostic tool for identifying issues like background bias.

Figure 19. TAM supports analyzing biased scenarios. The landbirds in these two images were mistakenly classified as waterbirds due to the synthesized water backgrounds. The TAM identified that this biased recognition arises from the influence of the background.

O. Extensive Failure Cases Study

An important function of TAM is to support developers in analyzing failure cases, thereby deepening their understanding of the model's shortcomings and enabling the development of better MLLMs. Generally, developers analyze errors by comparing the reply and answer, while TAM provides a clear visual view to understand them with more insights. As shown in Fig. 24, we list several failure cases with the error reason and corresponding analysis. We find that sometimes the model can successfully locate the target object, but lacks additional knowledge related to it thus replying falsely or refusing to answer (e.g., the train and cat in the left of Fig. 24). If the model focus on other regions out of the target, the answer is possibly to be wrong. For example, we the model looks at the wall, it replies "living wall", instead of the specific plant type the user asked for. Another error type is tolerable, that is synonyms, hypernyms, or hyponyms of answers (e.g., UK vs. England, fabric vs. nylon).

We further conduct case analysis on videos using Qwen2-VL-2B [51] in Fig. 25. Some error types are interesting. In the first row, we find the model already knows the object is a laptop when generating the token "pink". But it turns to the case sequentially. It indicates the answer is shifted by context (maybe trained with some corpus including "pink case"). Besides, the representation is not strong enough, and the model cannot divide the pattern of the pillow and doll in the third row. In the fifth row, the picture with a border is similar to a book, while it is attached to the wall. From this context, we can know this is a picture instead of a book, indicating the weak capacity to integrate context. For the last row, the attention is located on the hair, suggesting the model predicts the "washing" according to the hair, instead of the window. All these examples prove that TAM can provide more cues and insights to analyze failure cases.

P. Extensive Success VQA Examples

In addition to the failure cases illustrated in Fig. 24, we present extensive success Visual Question Answering (VQA) examples in Fig. 26. These visualization results indicate that the Token Activation Model (TAM) is applicable not only to caption-based datasets but also to VQA datasets, such as QK-VQA [37]. From the figure, we observe that certain images are well-aligned with the generated tokens, which include objects, actions, texts, and patterns (e.g., the Qantas logo), thereby facilitating accurate predictions. However, some cases are not primarily object-determined; they rely heavily on textual cues, as seen with terms like "commercial" and "cross" in the last row. This analysis allows us to discern the sources of predictions based on activation levels: higher responses indicate strong visual relevance, while lower responses suggest a greater reliance on textual information.

Q. Examples about Video Visualization

Video modality is a crucial input type for MLLMs; however, it has seldom been studied in the explainability aspect. We compare our TAM with conventional methods [45, 56], as illustrated in Fig. 27, using Qwen2-VL-2B [51] on the STAR dataset [52] for video understanding. It is evident that TAM produces significantly clearer video visualization results compared to CAM [56] and Grad-CAM [45], both of which are well-established methods, as shown in Table 3. Specifically, TAM effectively reduces redundant activations and minimizes noise, allowing users to concentrate on target objects and observe the raw video more clearly. Additionally, we provide case studies in Fig. 25 for video error analysis.

R. Corner Case About Reasoning

TAM serves as a valuable tool for analyzing the visual reasoning processes of MLLM. In Fig. 28, we present a corner case of visual reasoning and analyze it using TAM. We find that both Qwen2-VL-7B [51] and InternVL2_5-8B [16] provided incorrect answers in this case. TAM reveals that the primary issue arises from incomplete recognition of all arrows in the problem, with missing arrows leading to subsequent reasoning errors. Furthermore, the reasoning capability of Qwen2-VL-7B is weaker than that of InternVL2_5-8B in this instance, as evidenced by inconsistent context understanding and very low text activations. In contrast, while InternVL2_5-8B exhibited stronger textual responses and correct logical reasoning, it compromised visual perception and interpretability, ultimately resulting in an incorrect answer as well. These analyses offer valuable insights for future model optimization and highlight the broad applicability of TAM.

S. TAM for Multi-image Conversation

Conventional models generally have a single input and a single output, whereas the characteristic of MLLM is that it supports multiple inputs and multiple token outputs. We provide examples of multi-image conversation in this chapter, as shown in Fig. 29. In the first example, Qwen2-VL-2B [51] can identify the main elements from four different images, accompanied by high-quality explanation results provided by TAM. In the second example, TAM explains from a visual perspective why Qwen2-VL-2B considers the fourth image the most interesting. TAM presents detailed activation maps, explaining specific focus points on attributes such as object (pug), artwork (The Scream), image style (cartoon), and painting style (anthropomorphic). These results demonstrate the broad applicability of TAM, which can support various new capabilities of MLLMs.

T. TAM for Multi-turn Conversation

TAM supports multi-turn conversation for MLLM, which is a new capability compared to conventional models. We first present a qualitative example in Fig. 30. Qwen2-VL-2B [51] can effectively generate the image description, and TAM provides accurate response maps for various attributes, such as objects, actions, and text. Subsequently, the user engaged in multi-turn conversation, inquiring about a fatter dog and the color of a chair. TAM effectively interpreted these fine-grained tokens, including positional information, adjectives, and colors. This example demonstrates TAM's broad applicability and offers strong interpretability analysis for new features like multi-turn dialogue in MLLM.

Additionally, we provide an analysis of a faulty example in Fig. 31, showing that TAM helps locate model errors and provides visual insight for developers. Although Qwen2-VL-2B can recognize why this image is distinctive and demonstrates strong interpretability for object tokens, it made errors in understanding speed and motion blur. Specifically, the taxi exhibited motion blur indicating higher speed, but it incorrectly identified it as an SUV. In the second round of dialogue, we speculated that it might not have recognized the blur, or it could have recognized the blur but failed to understand the relation between blur and speed. Thus, in the third round of dialogue, we asked which vehicle exhibited blur, and the clues provided by TAM indicated that the failure to recognize motion blur was the main reason for this faulty example. TAM supports multi-turn conversation, allowing for more detailed analysis of errors and offering developers precise reasons and a deeper understanding for issues.

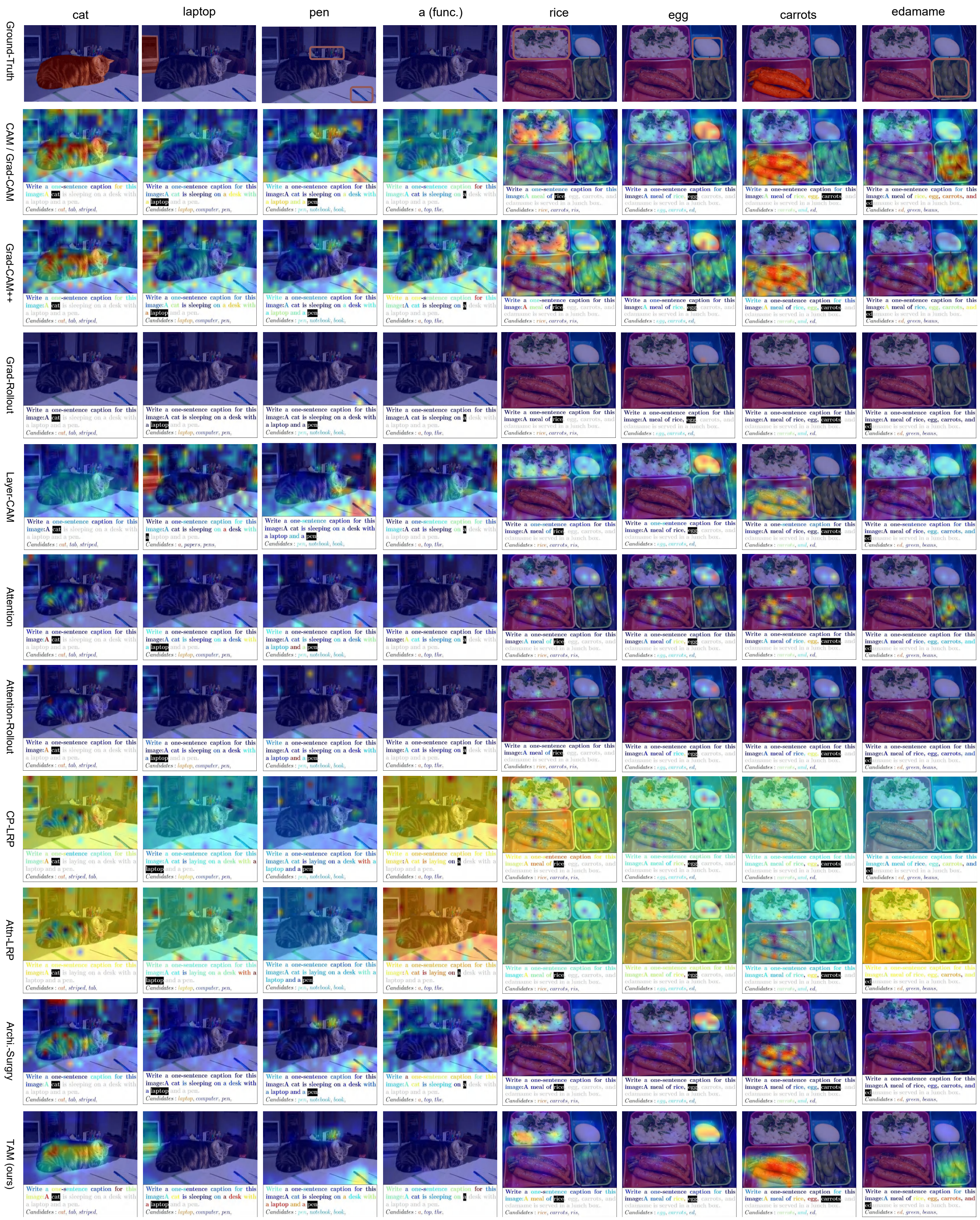


Figure 12. Visual comparison between our TAM and SoTA methods on the COCO Caption dataset [13] using the Qwen2-VL-2B [51] model. Objects without ground-truth are marked by red boxes. **TAM performs best beyond previous SoTA methods**. "CAM / Grad-CAM" indicates CAM [56] and Grad-CAM [45] are equivalent for MLLM as discussed in Supp. D.

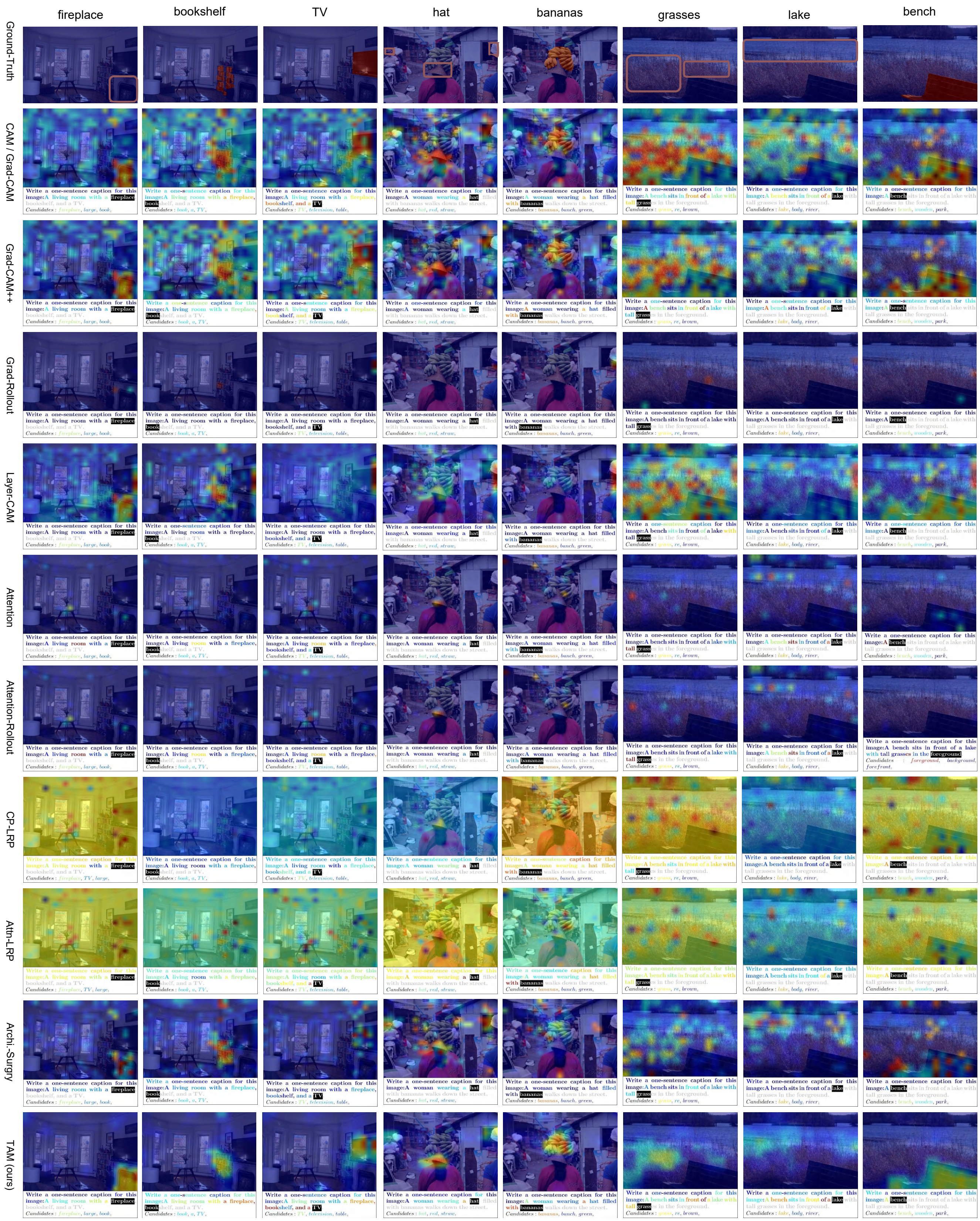


Figure 13. Visual comparison between our TAM and SoTA methods on the COCO Caption dataset [13] using the Qwen2-VL-2B [51] model. Objects without ground-truth are marked by red boxes. "func." indicates function words assigned as the background class. **TAM performs best beyond previous SoTA methods**.

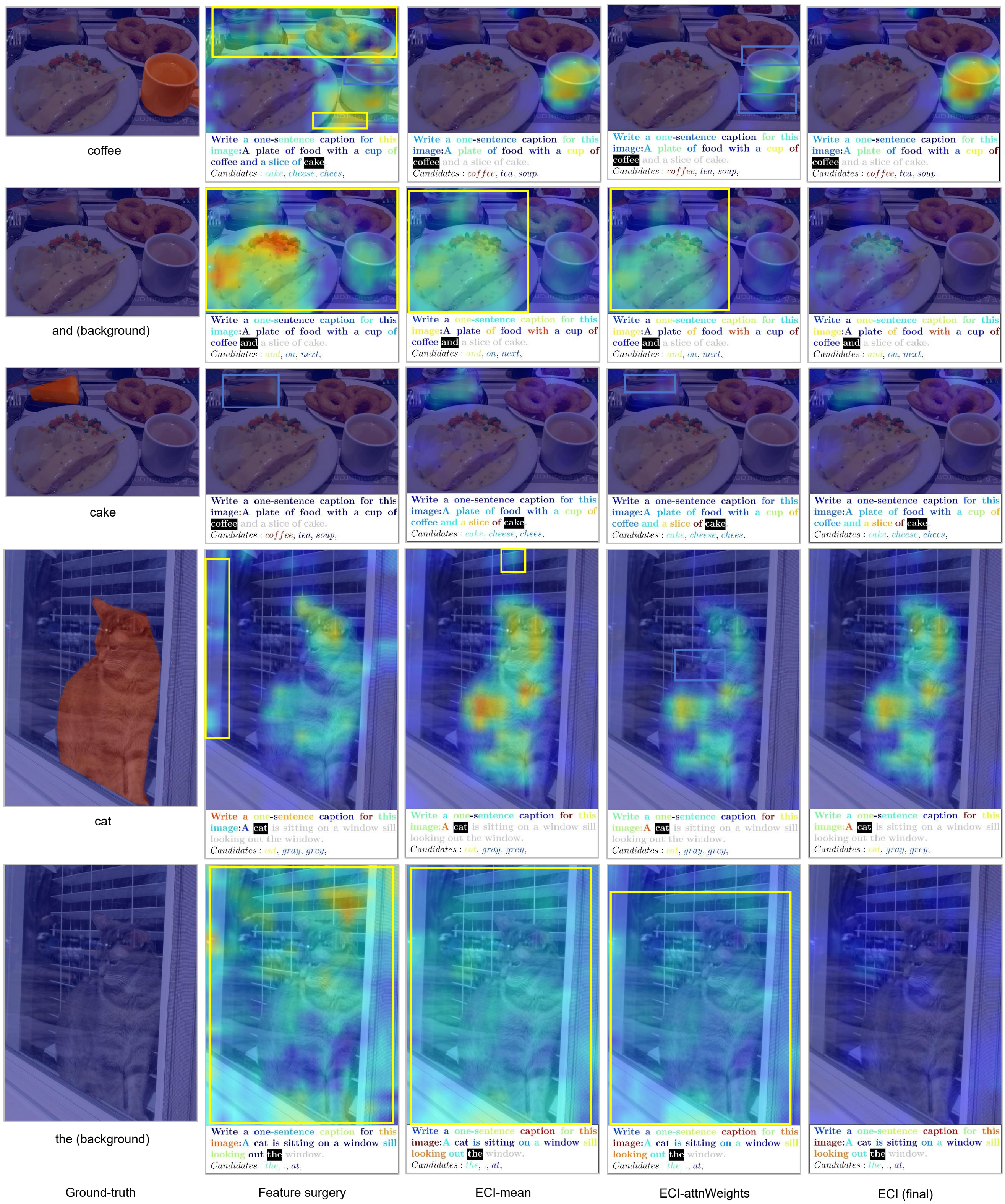
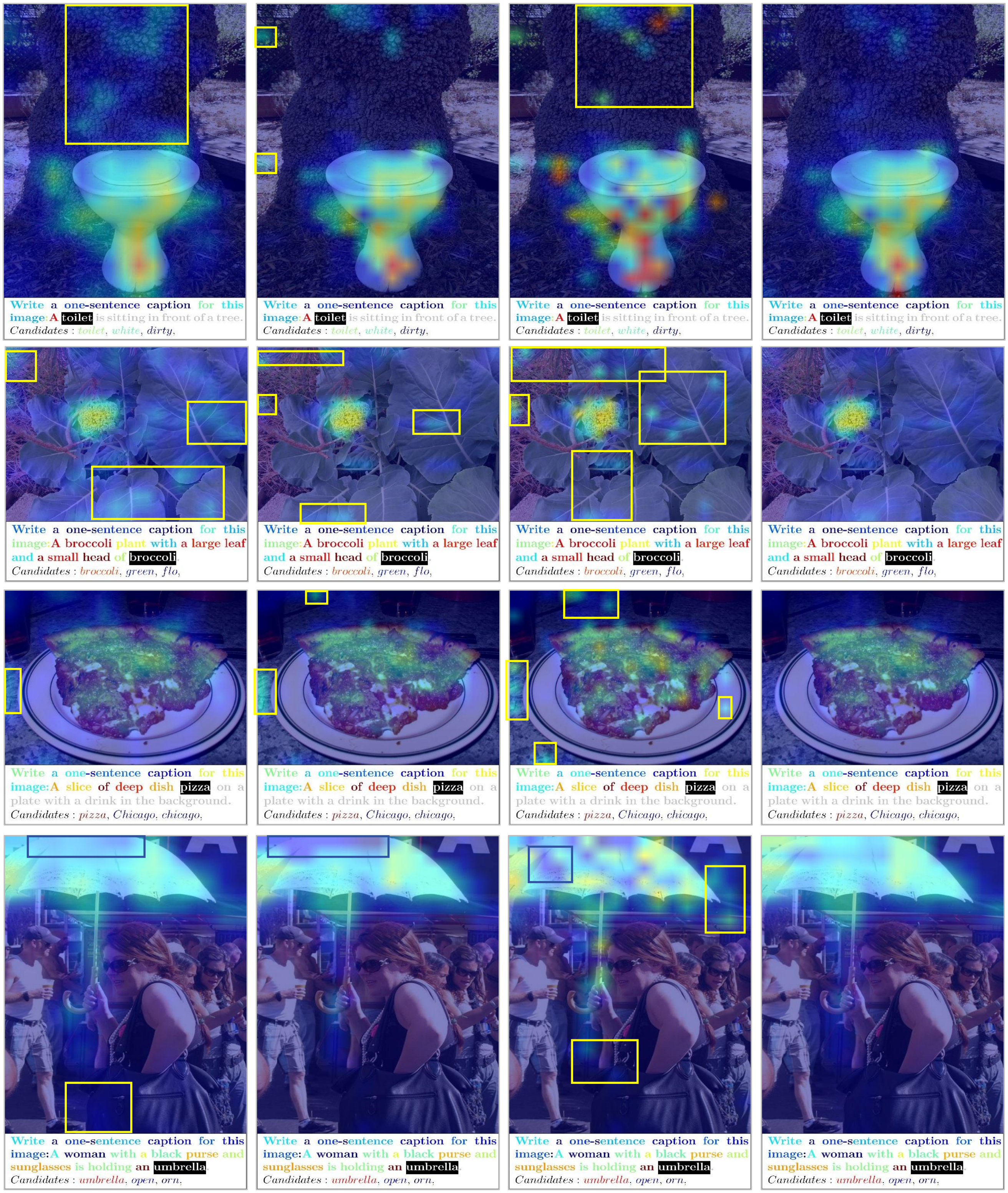


Figure 14. **The proposed estimated causal inference is well-designed beyond other methods and implementations**. The yellow boxes indicate correlated activations, and the blue boxes mean missed activations. Feature surgery [32] is designed for CLIP [41] to mitigate redundant features along the class dimension, while ECI-mean and ECI-attnWeights are other implementations of our estimated causal inference. The used model is Qwen2-VL-2B [51] on the COCO Caption dataset [13].



Gaussian filter | Median filter | Adaptive median filter | Rank Gaussian filter (ours)

Figure 15. **The proposed rank Gaussian filter is more effective than existing methods**. The yellow boxes indicate insufficient denoising, and the blue boxes mean over-denosing. The used model is Qwen2-VL-2B [51] on the COCO Caption dataset [13].

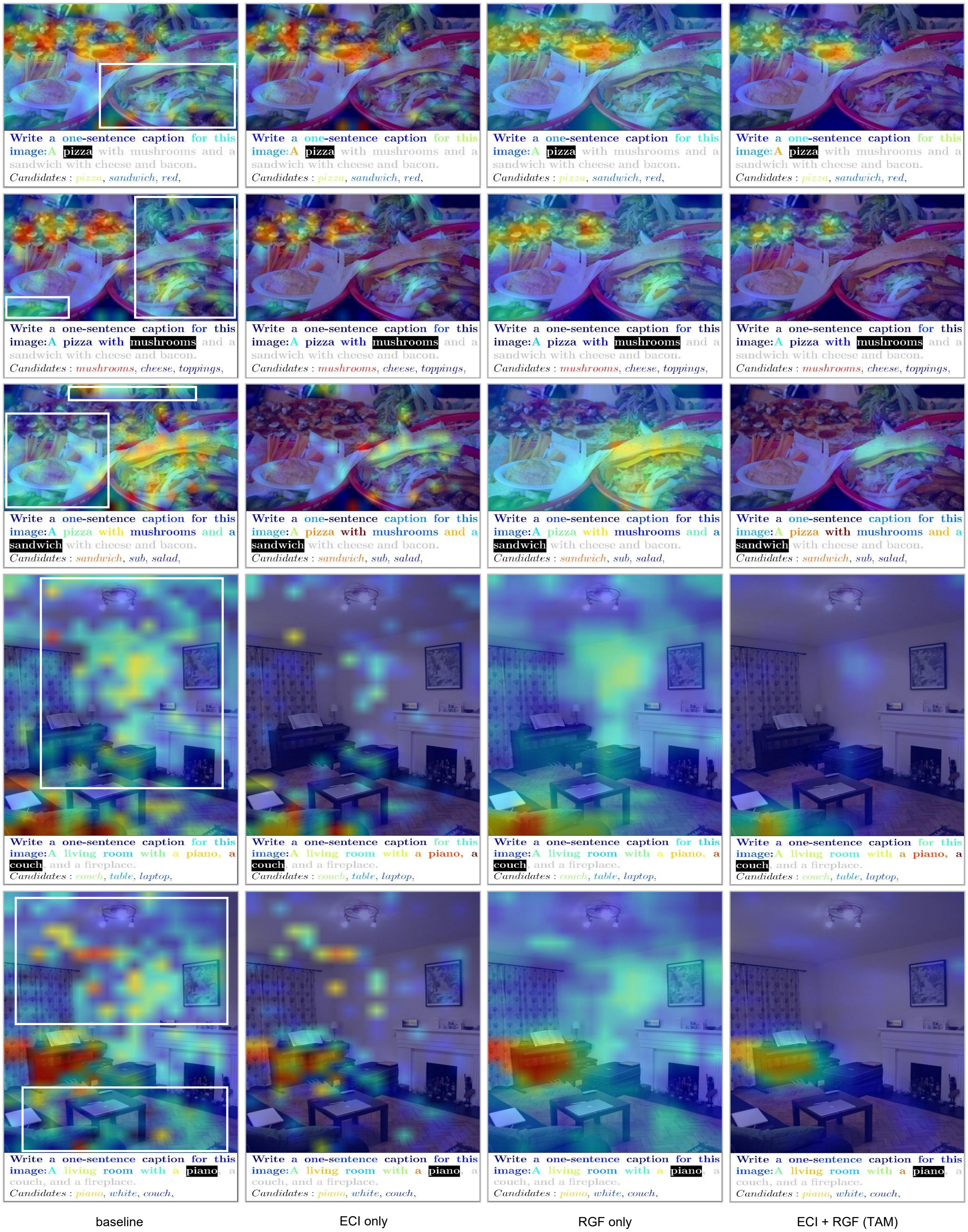


Figure 16. **TAM involves two mutually beneficial modules.** (ECI only) The estimated causal inference can mitigate most of the interference activations, while some stubborn noise remains. (RGF only) The rank Gaussian filter can remove these noises, while redundant activations are obvious. (ECI + RGF) TAM integrates the advantages of both and achieves good explanation results far beyond the baseline [45, 56]. The white boxes indicate correlated activations solved by our method. The used model is Qwen2-VL-2B [51] on the COCO Caption dataset [13].

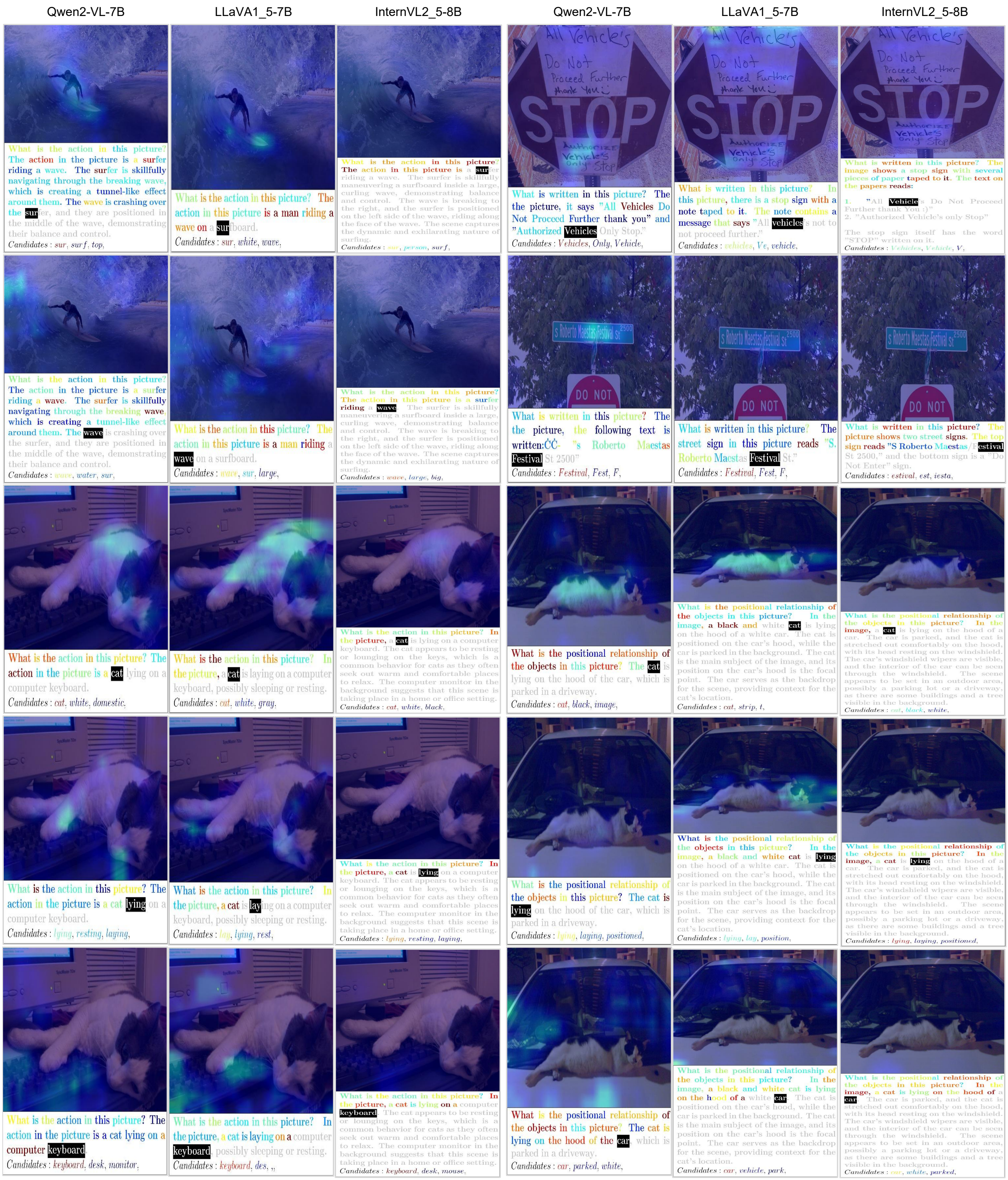


Figure 17. **TAM supports visual comparison among MLLMs about attributes**. Qwen2-VL-7B [51] presents good visual explainability beyond LLaVA1_5-7B [35] on texts. InternVL2_5-8B [16] focuses on textual content with more red texts and weaker visual activations.

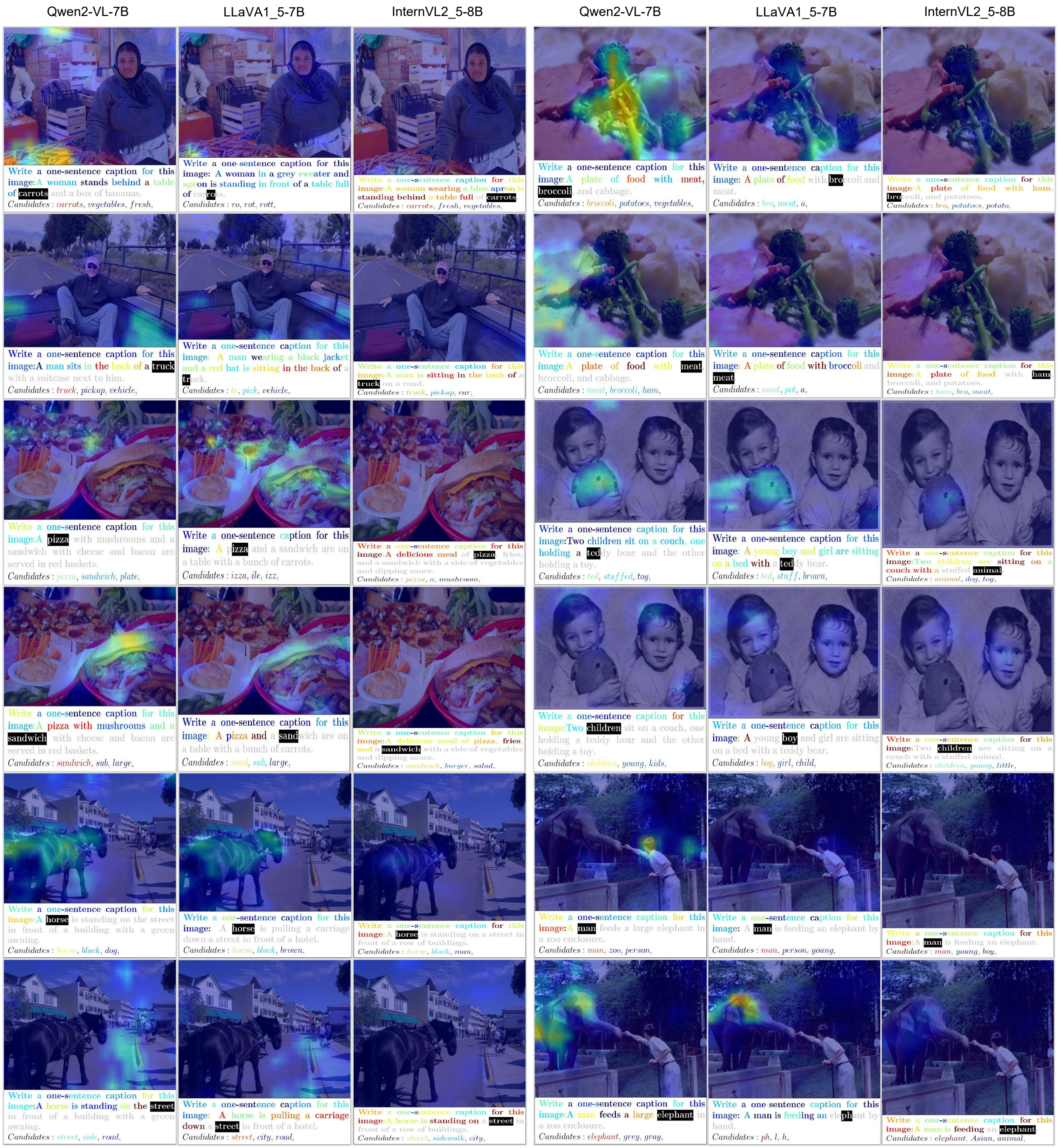
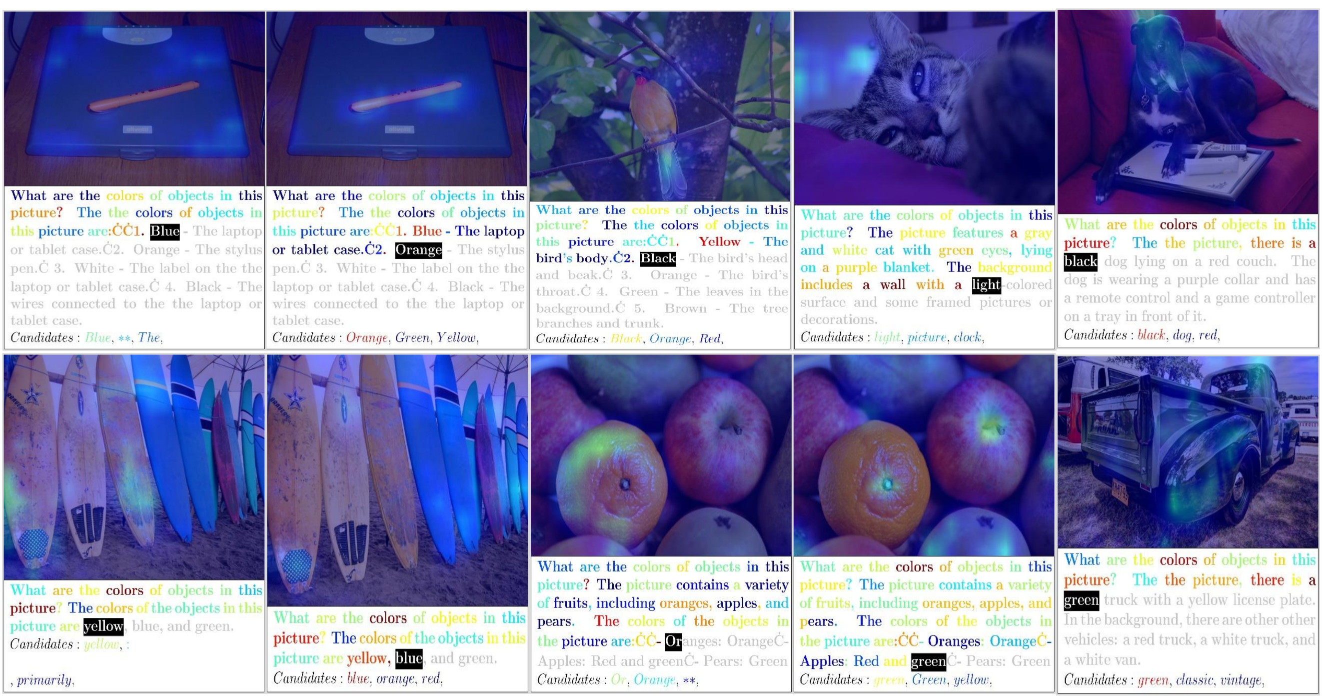


Figure 18. **TAM supports visual comparison among MLLMs about objects** on the COCO Caption [13] dataset. Qwen2-VL-7B [51] presents the best visual results with less correlation (e.g., pizza vs. sandwich in the third row) and shows a stronger activation degree.

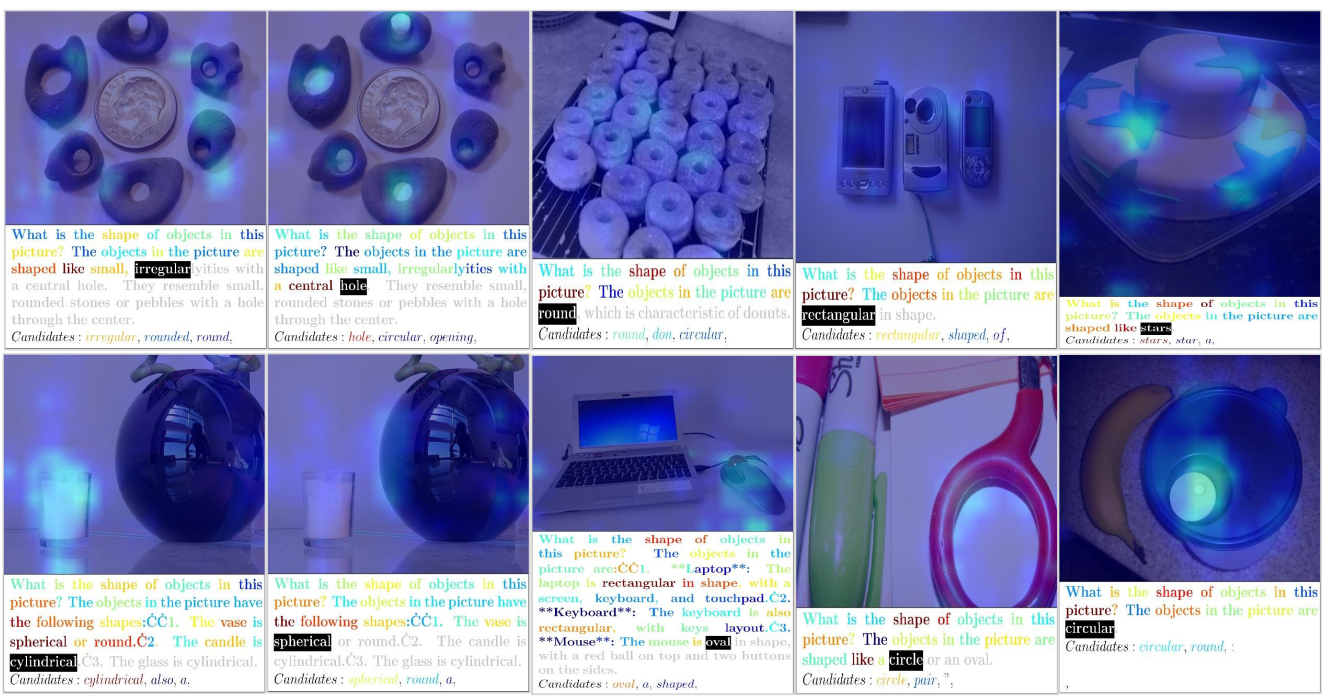
TAM to explain **actions**



TAM to explain **colors**

Figure 20. **TAM supports explaining attributes of MLLMs at high-quality** for the Qwen2-VL-7B [51] about action and colors.

TAM to explain **text**



TAM to explain **shape**

Figure 21. **TAM supports explaining attributes of MLLMs at high-quality** for the Qwen2-VL-7B [51] about text and shape.

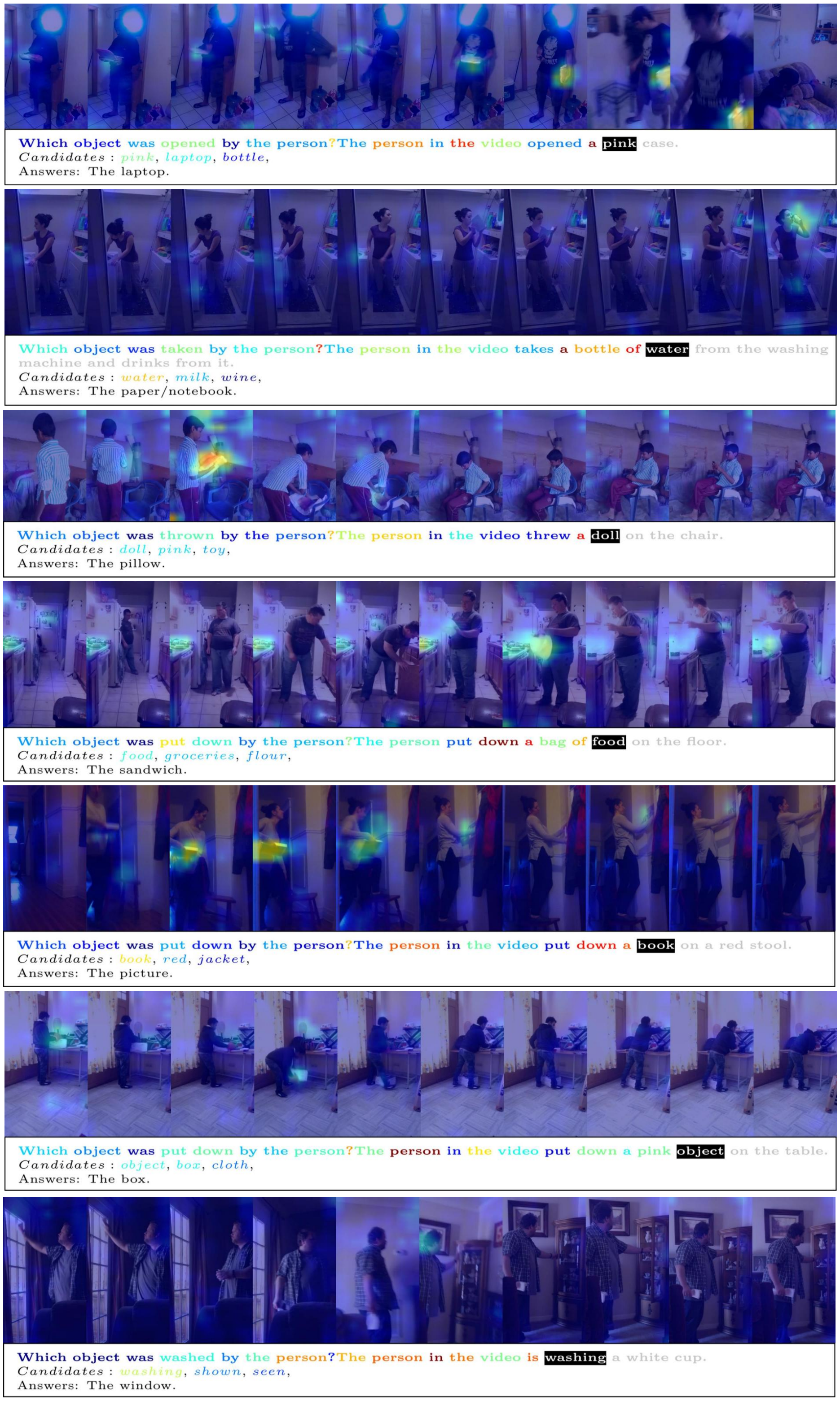
TAM to explain **location** in image

TAM to explain **location** in video

Figure 22. **TAM supports explaining attributes of MLLMs at high-quality for both images and videos.** The images are processed by the Qwen2-VL-7B [51] from the COCO Caption dataset [13] and we use the Qwen2-VL-2B for videos from the STAR dataset [52].

Figure 23. **TAM exceeds existing methods on attribute explanation.** The images are processed by the Qwen2-VL-7B [51] from the COCO Caption dataset [13] compared with the baseline. "CAM / Grad-CAM" indicates CAM [56] and Grad-CAM [45] are equivalent for MLLM, as discussed in Supp. D.

Figure 24. **TAM supports failure case analysis for deeper understanding** with error reason and analysis details using Qwen2-VL-2B [51] on the QK-VQA dataset [37].



Type: answer shift by context

Details: the second prediction of token “pink” is the right answer. But after generating this context, the final answer is shifted to case from laptop.

Type: incomplete answers

Details: the given answer is paper / notebook. But in the last frame, the women take a bottle of water, which is also right.

Type: weak representations

Details: the model locates the right object, but recognize the pillow as a doll, owing to weak representations that cannot divide them.

Type: hypernym of the answer

Details: the word food is hypernym of sandiwich. This answer is not wrong, but not specific.

Type: weak representations and context

Details: the picture / photo is similar to the book in apperance. But the book is not supposed to be attached on the wall. Here the context is not well considered.

Type: hypernym of the answer

Details: the word object is hypernym of box. This answer is not wrong, but not specific.

Type: weak action understanding

Details: the word “washing” is related to the hair accroding to the activations, indicating the missed wiping action.

Figure 25. **TAM supports failure case analysis for videos** with error reason and analysis details using Qwen2-VL-2B [51] on the STAR dataset [52].

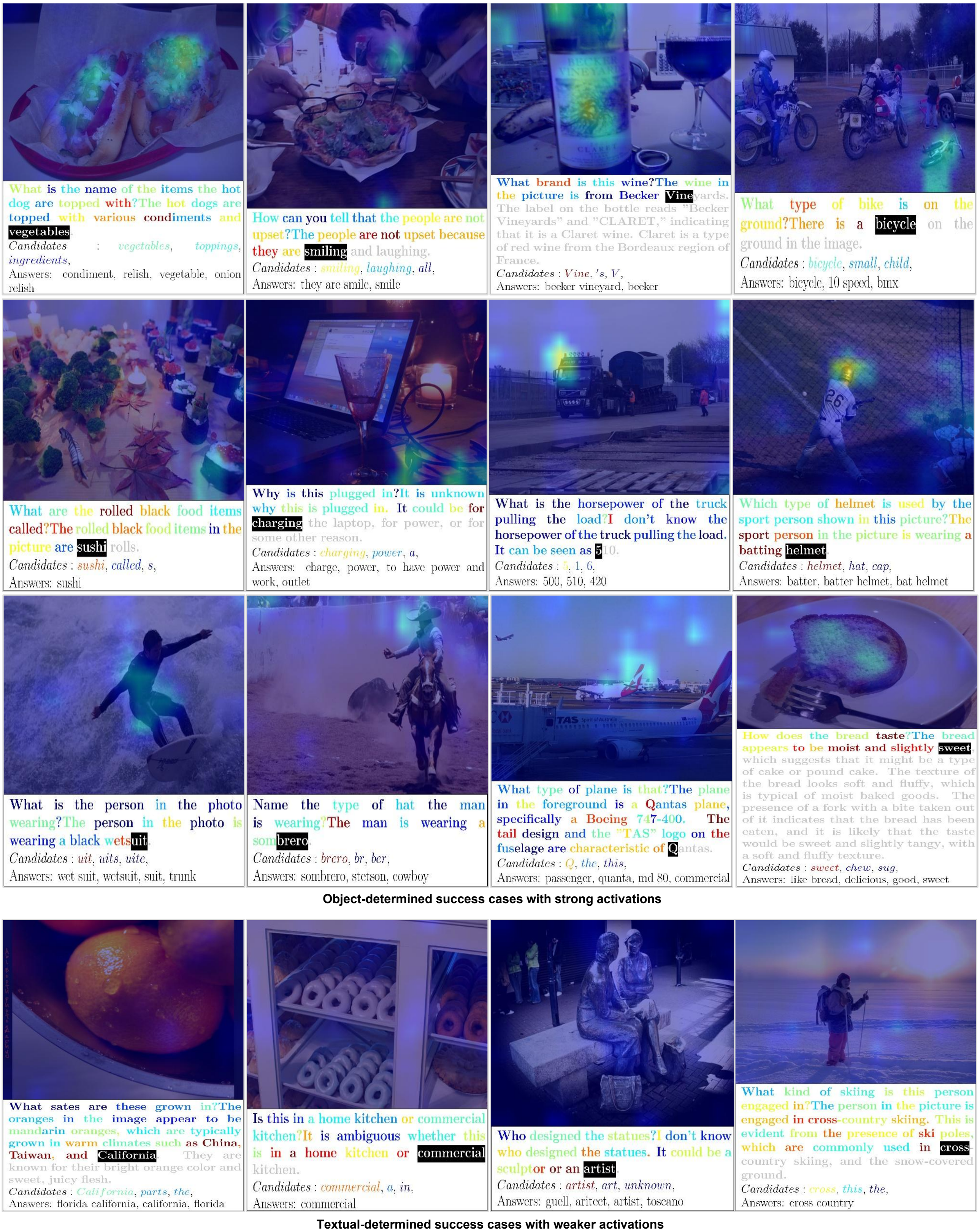


Figure 26. **TAM presents good visual explanation result for the VQA dataset** with extensive successful examples on the QK-VQA dataset [37] using Qwen2-VL-2B [51]. These cases are dependent on different information, divided into "Object-determined" type and "Textual-determined" type, with higher and lower activation degrees, respectively.

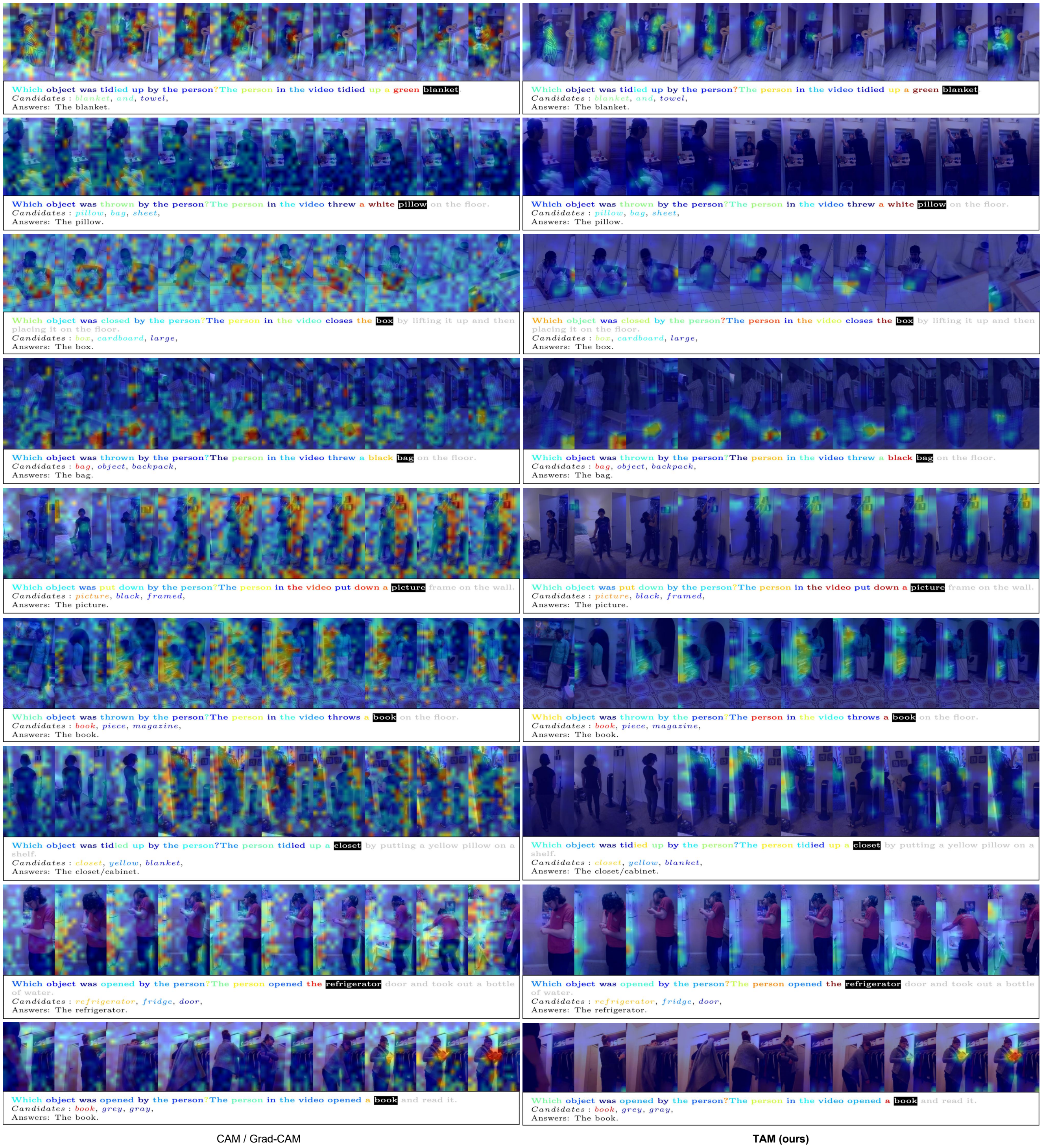
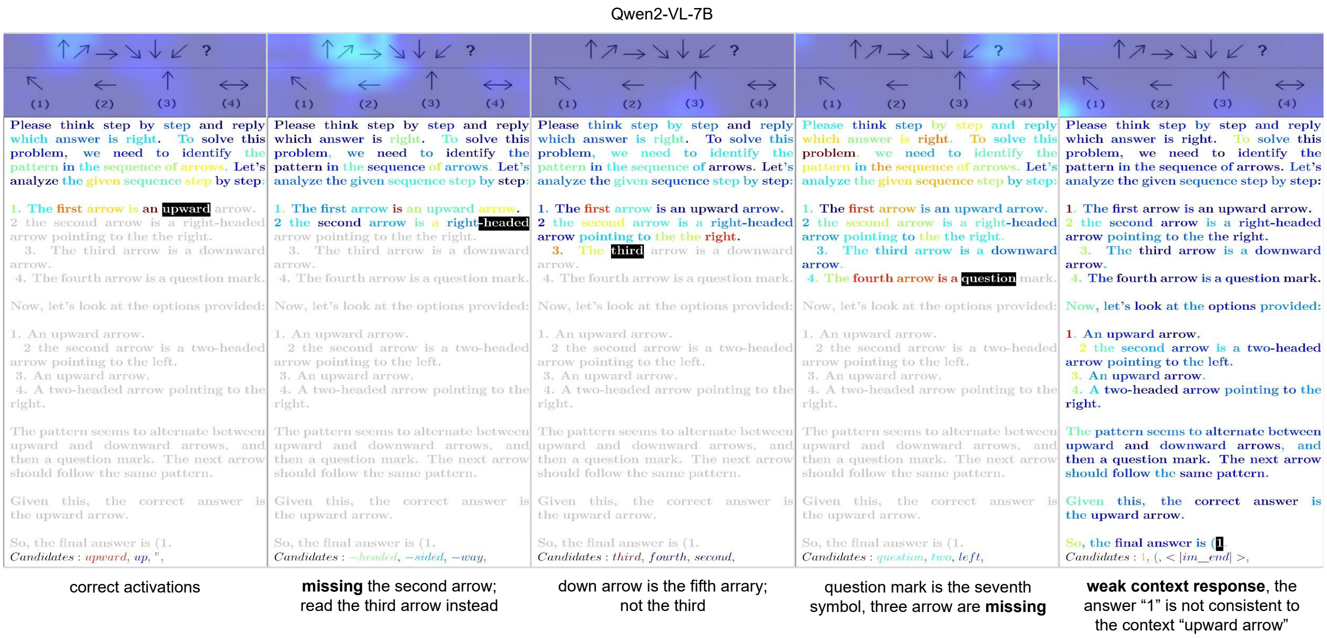


Figure 27. Comparison about video visualization between our TAM and CAM [56] / Grad-CAM [45] (equivalent to MLLM) on the STAR dataset [52] using Qwen2-VL-2B [51]. **TAM presents much clearer visualization results with fewer redundant activations and noises**.



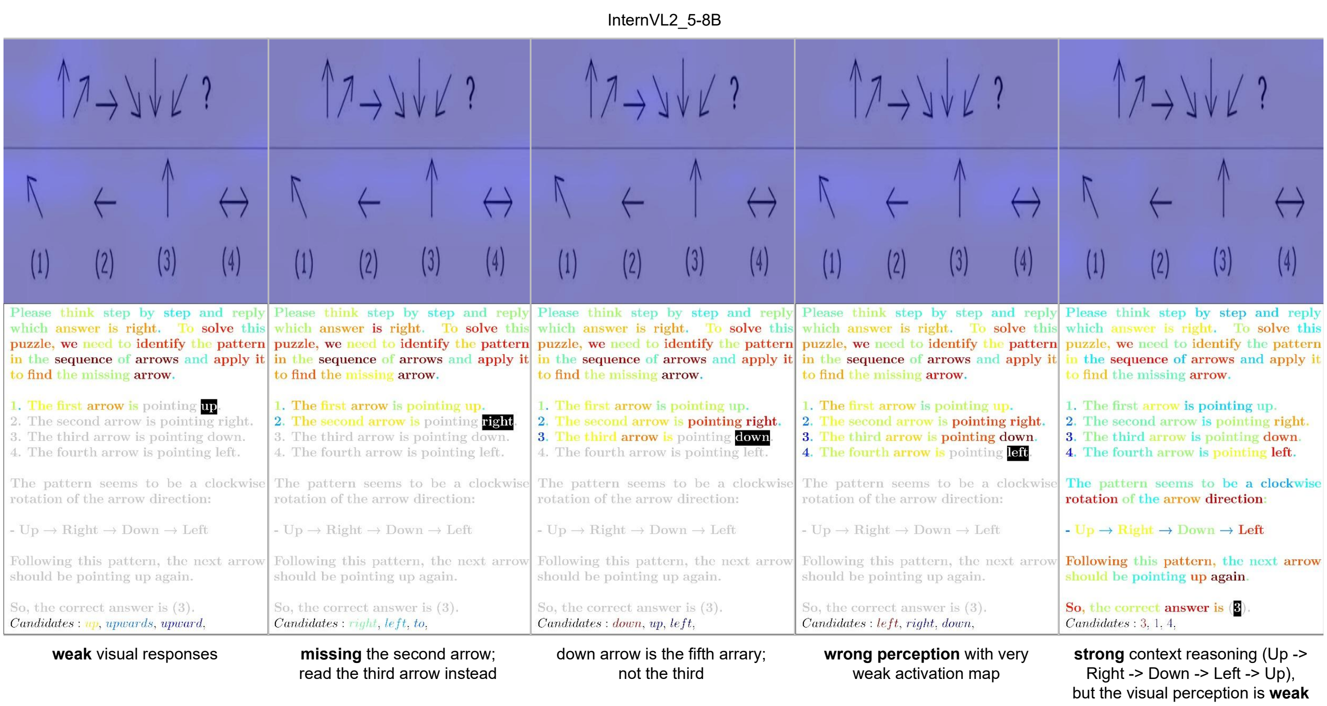


Figure 28. **Visual reasoning corner case analyzed using TAM**. The analysis reveals that both Qwen2-VL-7B [51] and InternVL2_5-8B [16] incorrectly answered the question due to incomplete recognition of all the arrows. Missing arrows led to erroneous reasoning. Qwen2-VL-7B demonstrated weaker reasoning capabilities compared to InternVL2_5-8B, exhibiting inconsistent context understanding and low text activations. In contrast, while InternVL2_5-8B provided a stronger text response with correct reasoning logic, its visual perception and activation degree are weaker.

Figure 29. **TAM supports multi-image conversation of MLLM showing wide applicability.** It generates high-quality activation maps for four input images in the first case using the Qwen2-VL-2B [51]. In the second case, TAM explains why the model regards the last image as the most interesting one, providing visual cues for key tokens. These results showcase the model's effectiveness in multi-image scenarios, highlighting its versatility in handling complex visual data.

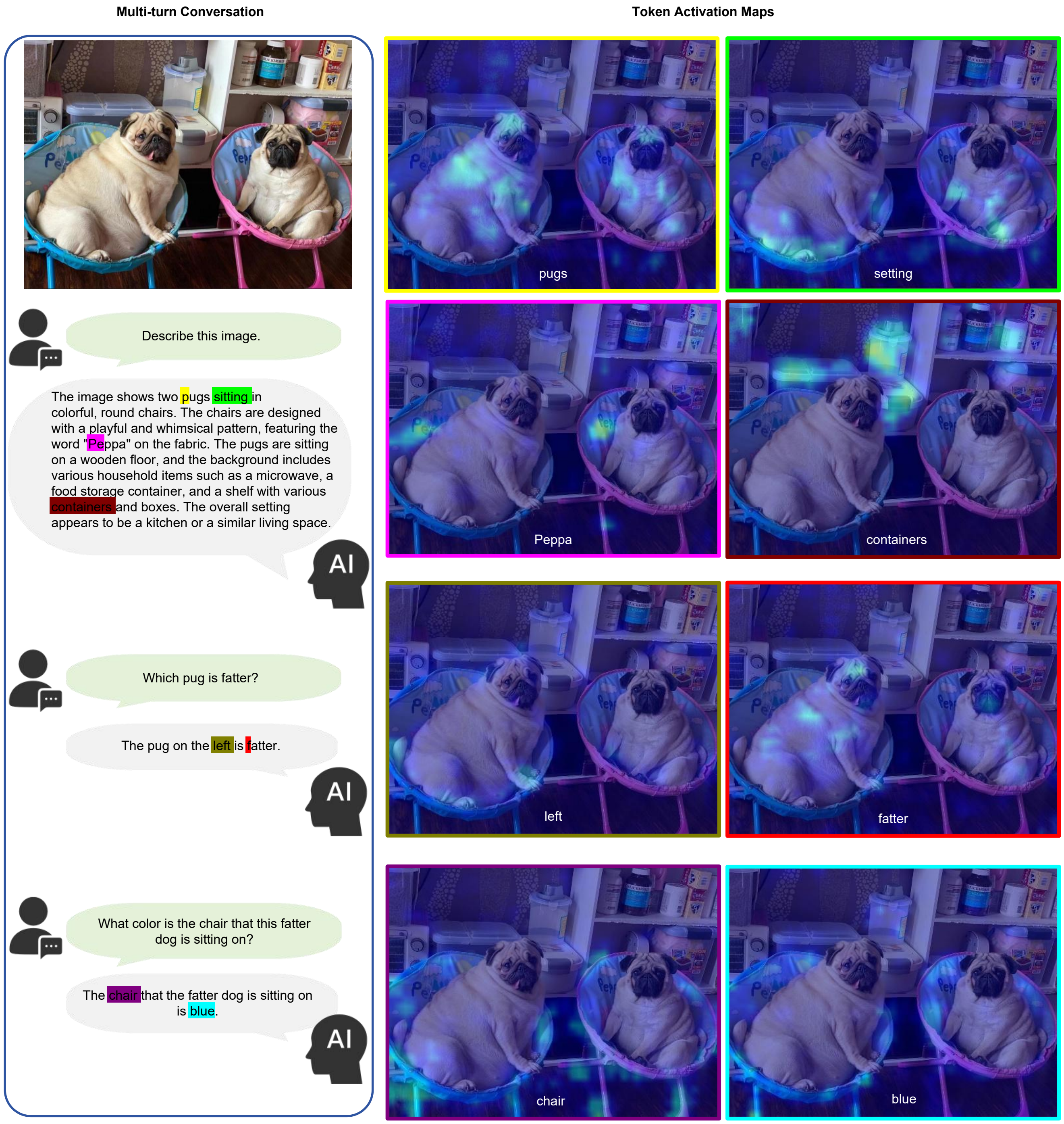


Figure 30. **TAM supports multi-turn conversation of MLLM for diverse attributes**. TAM presents high-quality visual explanation results on Qwen2-VL-2B [51] regarding attributes such as objects, actions, and text in the first round. Then, the user inquires about the fatter dog and the color of a chair in the second and third rounds, respectively. Activation maps suggest TAM is capable of explaining fine-grained tokens like positional information, adjectives, and colors in multi-turn conversation.

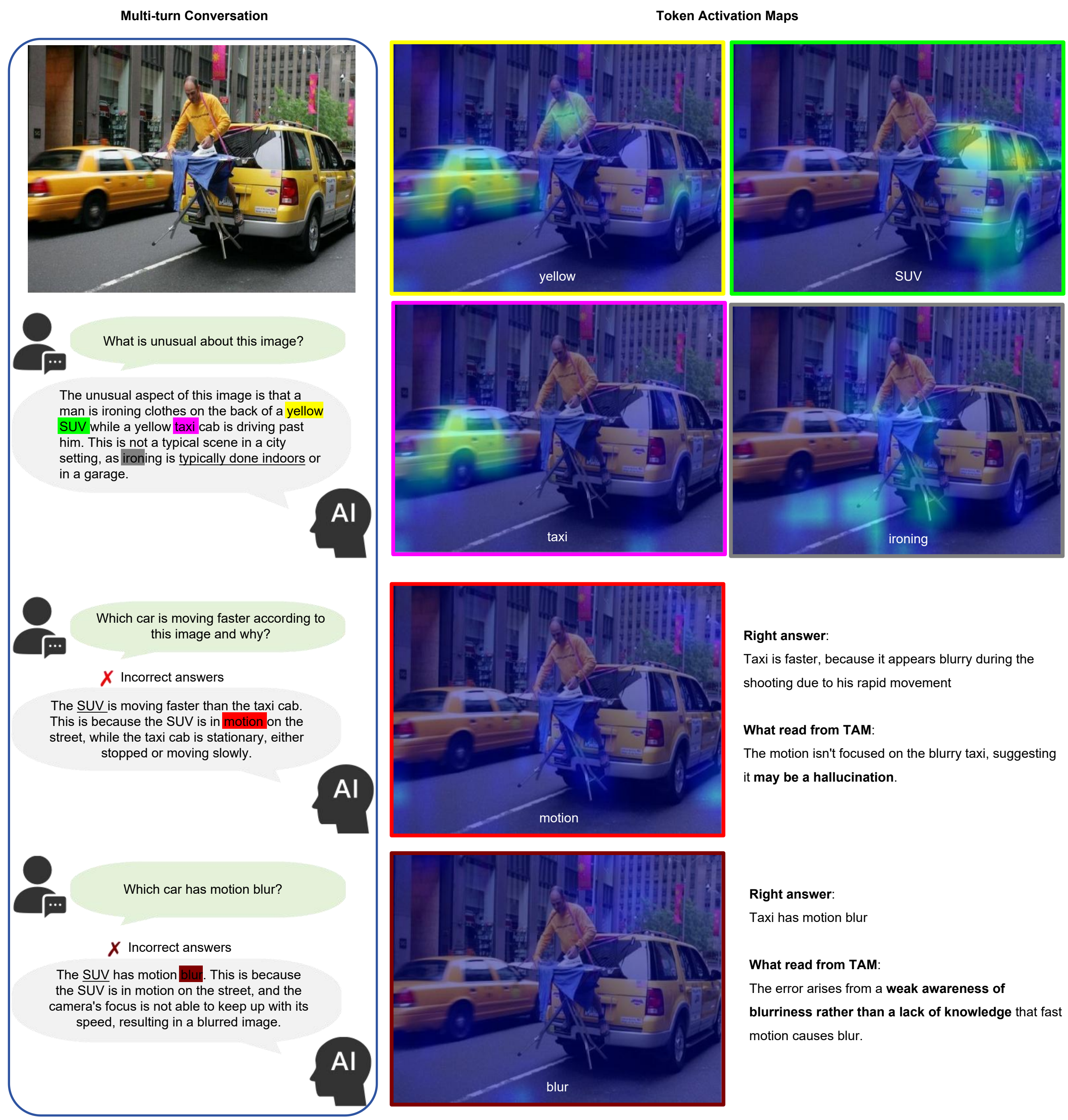


Figure 31. **TAM enables failure case analysis in multi-turn conversation**. Although Qwen2-VL-2B [51] well recognizes objects with good explanation results in the first round chat, it fails to identify motion blur related to speed and mistakenly regards the SUV as the faster car. The clues provided by TAM reveal that the failure to recognize motion blur is the primary reason for this error, highlighting TAM's effectiveness in supporting detailed error analysis from multi-turn conversation.